World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:18, No:12, 2024

Leveraging SHAP Values for Effective Feature Selection in Peptide Identification

Authors : Sharon Li, Zhonghang Xia

Abstract : Post-database search is an essential phase in peptide identification using tandem mass spectrometry (MS/MS) to refine peptide-spectrum matches (PSMs) produced by database search engines. These engines frequently face difficulty differentiating between correct and incorrect peptide assignments. Despite advances in statistical and machine learning methods aimed at improving the accuracy of peptide identification, challenges remain in selecting critical features for these models. In this study, two machine learning models—a random forest tree and a support vector machine—were applied to three datasets to enhance PSMs. SHAP values were utilized to determine the significance of each feature within the models. The experimental results indicate that the random forest model consistently outperformed the SVM across all datasets. Further analysis of SHAP values revealed that the importance of features varies depending on the dataset, indicating that a feature's role in model predictions can differ significantly. This variability in feature selection can lead to substantial differences in model performance, with false discovery rate (FDR) differences exceeding 50% between different feature combinations. Through SHAP value analysis, the most effective feature combinations were identified, significantly enhancing model performance.

Keywords : peptide identification, SHAP value, feature selection, random forest tree, support vector machine
Conference Title : ICDMKE 2024 : International Conference on Data Mining and Knowledge Engineering
Conference Location : New York, United States
Conference Dates : December 09-10, 2024